ULMER TEXTBANK

Datenbank als Mittel der Psychotherapieforschung.

H. Kächele und E. Mergenthaler

Vortrag am SFB 115 der Universität Hamburg, Universitäts-Krankenhaus Eppendorf

Januar 1983

Abteilung Psychotherapie
Universität Ulm
Am Hochsträß 8

D 7900 Ulm

SFB 129 - Projekt B2

Einleitung

In der ersten Auflage des Handbuches für Psychotherapieforschung aus dem Jahre 1971 stellten Luborsky und Spence fest, daß die psychoanalytische Forschung durch die Dürftigkeit primärer Daten, also jener Daten, die in der aktuellen analytischen Situation erhoben werden, erschwert sei. Auch in der zweiten Auflage dieses Handbuches im Jahre 1978 beklagen sie nach wie vor den Mangel an guten Primärdaten. Allerdings können sie bereits über elf vorwiegend in den Vereinigten Staaten verfügbare Datenbanken mit psychoanalytischen Behandlungsprotokollen berichten. Wir werden Ihnen heute über Aufbau, Umfang und Ziele einer dieser elf Datenbanken, der ULMER TEXTBANK berichten und einige der sich daraus ergebenden Forschungsmöglichkeiten aufzeigen. Wie wir erst kürzlich im persönlichen Gespräch mit Luborsky anläßlich seines Besuches in Ulm feststellen konnten, stellt dieses Vorhaben bis heute eine Besonderheit dar. Dies hat Luborsky unter anderem veranlaßt, selbst 80 Verbatim-Protokolle in die Textbank einzubringen.

Der Weg zur Textbank

Der Grundstein zur ULMER TEXTBANK wurde vor rund zehn Jahren gelegt, als an der Abteilung für Psychotherapie der Universität Ulm die ersten Tonbandaufnahmen von psychotherapeutischen Behandlungen durchgeführt wurden. Nach anfänglichen manuellen Auswertungen an den normalschriftlich transkribierten Gesprächen werden nunmehr seit sieben Jahren die Texte in maschinenlesbarer Schrift transkribiert und mit Hilfe von computerunterstützten Methoden ausgewertet. Freilich häuften sich zunächst die Probleme bei diesen Arbeiten, da die seinerzeit verfügbaren Programme für derartig große Textmengen, wie sie therapeutische Gespräche darstellen, nicht ausgelegt waren. Das EVA-System (GRÜNZIG? HOLZSCHECK und KÄCHELE 1976, MERGENTHALER und BÜSCHER 1976) eine Programmsammlung zur computerunterstützten Analyse von Texten und seinerzeit vom sozialwissenschaftlichen Seminar der Universität Hamburg übernommen, war "von Haus aus" zur Analyse von Schlagzeilen und Zeitungsannoncen ausgelegt. Erst eine gründliche Überarbeitung, die zu der bis heute verwendeten Ulmer Version führte, gewährleistete seinen problemfreien Einsatz. Damit war aber auch der Weg frei für fortführende Untersuchungen. In der Folge wurden Verbatim-Protokolle weiterer psychoanalytischer Behandlungen erstellt, Erstinterviews wurden verschriftet. Motiviert durch vielerlei Fragestellungen wuchs schließlich ein Fundus an Gesprächsprotokollen heran, der deutlich werden ließ, daß er alsbald nicht mehr überschaubar sein würde. Wie konnte sichergestellt werden, daß dieses kostbare Material (die Verschriftung einer Behandlungsstunde benötigt bis zu 35 Stunden) jederzeit auffindbar, gezielt auswählbar und ökonomisch verarbeitbar bleibt? Die zur Verfügung stehende Rechnertechnologie bot und bietet bis heute lediglich die Möglichkeit, Texte als Dateien auf externen Datenträgern abzuspeichern. Um für eine geplante Untersuchung ein Textkorpus anhand vorgegebener Kriterien zusammenzustellen, bedarf es damit nicht nur eines guten Erinnerungsvermögens für Dateinamen, Datenträger und zugeordneten Inhalt, sondern auch umfangreicher EDV-Kenntnisse, um über Sprachmittel des Betriebssystems, der Job Control Language,

die gewünschten Textmengen zusammenzubringen. Es lag nahe, nach Lösungen zu suchen, die diese Arbeit ebenfalls dem Computer übertragen, nach Werkzeugen, die dem an der Psychotherapieforschung Interessierten ohne großen Lernaufwand zur Verfügung stehen, und deren Leistungen er über eine ihm verständliche und anwendungsorientierte Kommandosprache steuern und erhalten kann.

Die Verwirklichung solcher Vorhaben ist Aufgabe der angewandten Informatik. Unter dem Oberbegriff der computerunterstützten Informationssysteme lassen sich eine Reihe von Problemlösungen finden, die auch für derartige Entwicklungen typisch sind. In diesem Sinne wurde nun im Frühjahr 1979 das Projekt der ULMER TEXTBANK konzipiert mit dem Ziel, ein computerunterstütztes Informationssystem für Texte zu erstellen und eine systematische Sammlung von Texten aus der psychotherapeutischen Situation anzulegen. Seit Januar 1980 konnte dieses Vorhaben mit Unterstützung der Deutschen Forschungsgemeinschaft im SFB 129 als Projekt schrittweise umgesetzt werden.

Das TEXT BASE MANAGEMENT - System

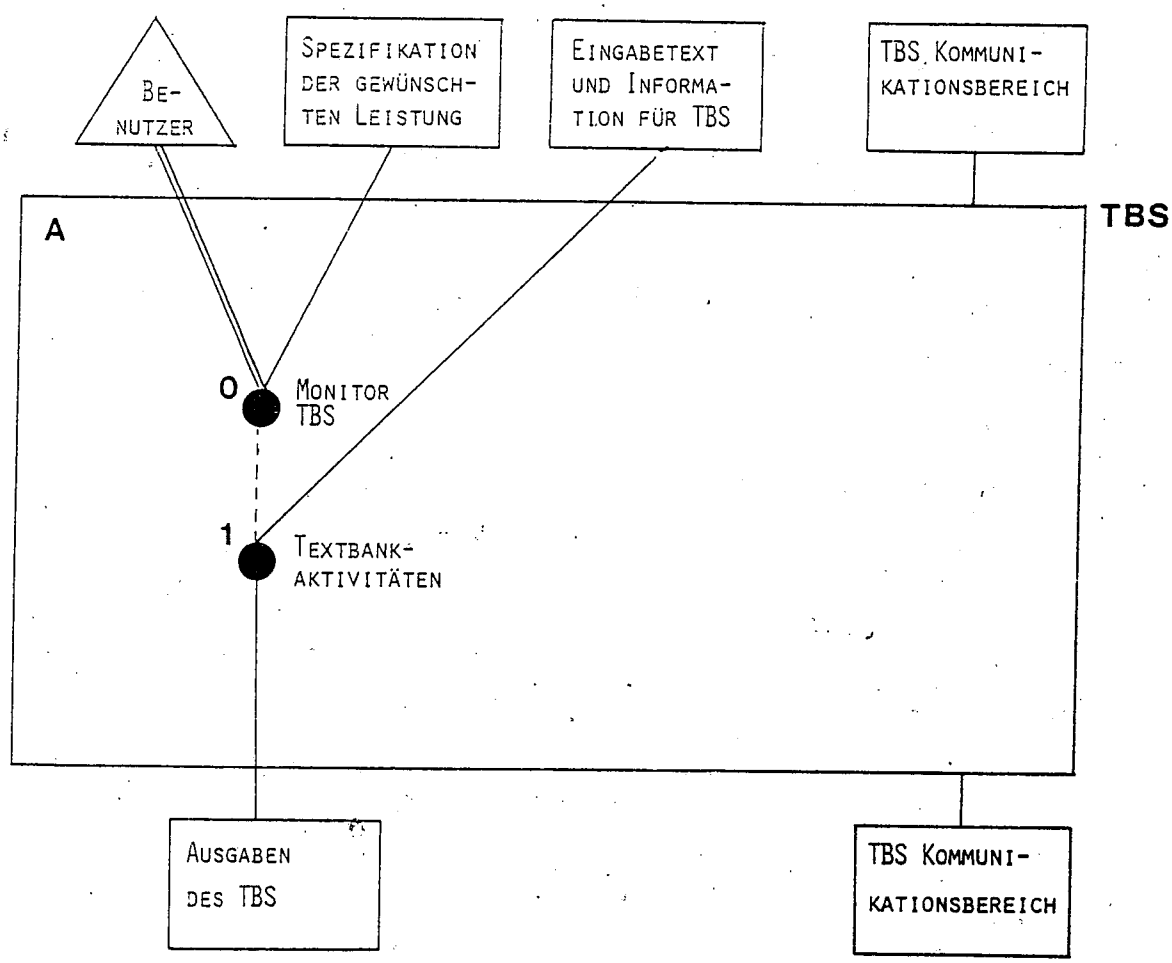
Unter dem Begriff des TEXT BASE MANAGEMENT - Systems (TBS) wird ein Computerinformationssystem verstanden, das drei unterschiedliche Aufgaben zu einem Leistungsangebot integriert:

Verwaltung beliebig vieler Texteinheiten,
Ermittlung textimmanenter Kenndaten und
Information über vorhandene Texte.

Es fällt damit in die Klasse der Datenbanksysteme mit der Möglichkeit zum Fakten- wie auch zum Dokumenten-Retrieval. Entsprechend umfaßt die Systemarchitektur Bausteine zur

Handhabung der Texte, zur
Analyse der Texte und zur
Verwaltung aller relevanten Informationen.

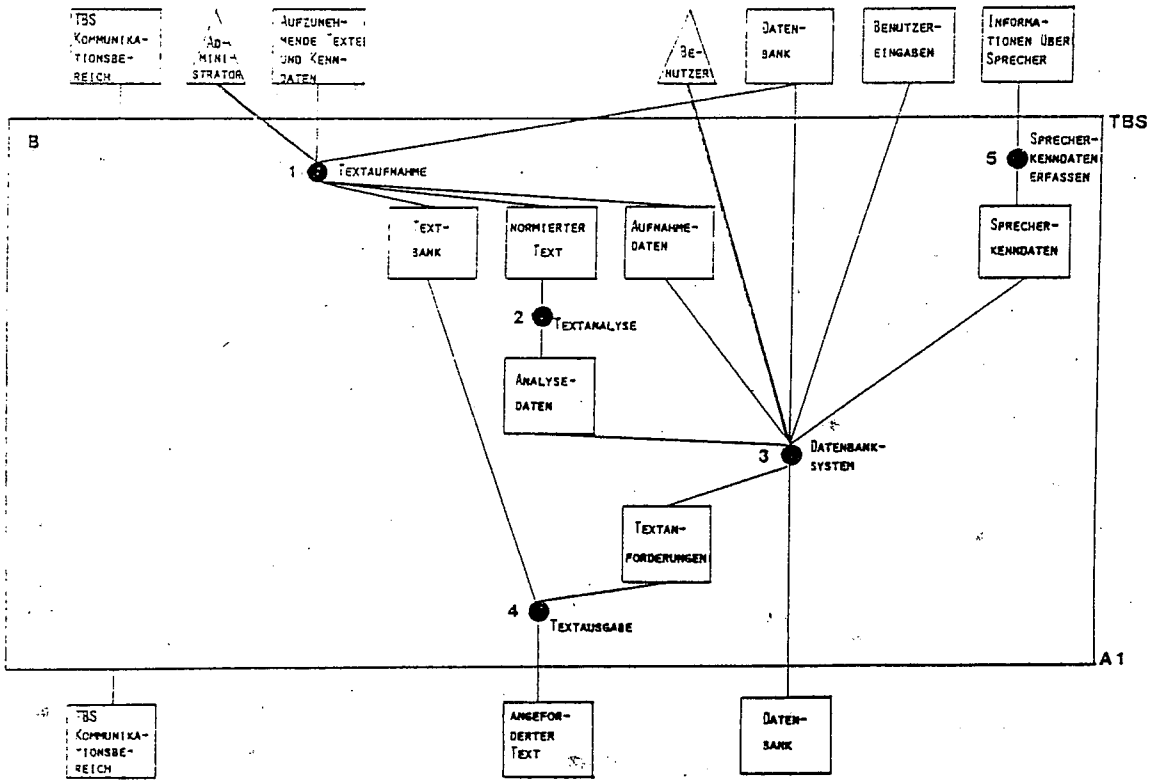
Die Verbindung zum Benutzer wird über ein Steuerprogramm, dem Monitor, hergestellt, der gleichzeitig die einzelnen Systemaktivitäten koordiniert und als Bindeglied zu anderen Softwarepaketen wie beispielsweise COCOA, einem Programm zur Bildung einer Konkordanz, dient. Das hierzu erforderliche Schnittstellenkonzept ist variabel und offen für maßgeschneiderte ad-hoc-Software. (Bild 1)
Die einzelnen Textbankaktivitäten werden als "Bausteine" des TBS nachfolgend kurz beschrieben.



Der Textbaustein

Der Textbaustein stellt ein Teilsystem zur Aufnahme und Ausgabe verschiedener Textformate dar (B1 und B4 in Bild 2). Hierzu sind eine Reihe von Schnittstellen definiert, von

denen die wichtigsten nachfolgend aufgeführt werden.



Bei der Aufnahme von Texten kann grundsätzlich zwischen der

- Off-line und der
- On-line Texterfassung

unterschieden werden. Die Off-line-Verarbeitung ist dadurch gekennzeichnet, daß von einer Schreibkraft in maschinenlesbarer Schrift (z. Zt. OCR-A) Belege erstellt werden, die über einen programmierbaren Belegleser verarbeitet und dem Textbaustein zugeführt werden. Bei der On-line-Verarbeitung gibt die Schreibkraft unmittelbar am Datensichtgerät den Text in den Rechner, in der Regel ein Mikrocomputer, ein (siehe dazu Schmucker 1981). Sie wird dabei weitgehend (etwa bei der Erkennung von Tippfehlern) von einem Erfassungsprogramm unterstützt.

Bei beiden Verarbeitungsformen ist der Text durch zusätzliche Information über die Sprecher und die Sprechsituation angereichert. Alle Eingaben werden eingehenden Plausibilitätskontrollen unterzogen. Die Einhaltung der Transkriptionsregeln wird überwacht. Im Hinblick auf die Anforderungen des Datenschutzes durchlaufen die Texte bereits bei der Aufnahme ein Anonymisierungsverfahren. Durch die Programme des Textbausteins wird nun der aufzunehmende Text in eine systeminterne Form, deren Struktur sich an den Erfordernissen einer optimalen Weiterverarbeitung orientiert, transformiert und in der 'TEXTBANK' (siehe Bild 2) abgelegt. Gleichzeitig werden die mit dem Text verbundenen Informationen zu den Sprechern und der Sprechsituation ermittelt und gegebenenfalls, beim Fehlen obligater Daten, vom Textbankadministrator erfragt und vervollständigt. Die Primärdaten werden zusammen mit den Kenndaten der Datenträger (Magnetband, Magnetplatte, Disketten) als 'Aufnahmedaten' (siehe Bild 2) systemintern zur Weiterverarbeitung bereitgestellt. Dies gilt auch für den 'normierten Text' (siehe Bild 2), einer Kopie des aufgenommenen Textes.

Bei der Ausgabe eines Textes werden systemintern 'Textanforderungen' (siehe Bild 2) erwartet mit Informationen darüber, auf welchen Datenträger die auszugebenden Texteinheiten zu finden sind, und in welcher Form sie ausgegeben werden sollen. Unterschieden wird eine Druckschnittstelle und eine Datenschnittstelle. Der Druck von Texten wird über ein gesondertes Layout-Programm verwirklicht. Eine Vielzahl von Gestaltungsmöglichkeiten durch den Benutzer sind vorgesehen. An der Datenschnittstelle werden die auszugebenden Texte in normierter Form zur Verfügung gestellt, wie sie zur weiteren linguistischen Datenverarbeitung im Rahmen des Gesamtsystems oder in Verbindung mit anderen Anwendersystemen geeignet ist.

Der Analysebaustein

Der Analysebaustein (B2 in Bild 2) ist ein Teilsystem zur Ermittlung formaler, grammatischer und inhaltlicher Merkmale von Texten. Er stellt den Rahmen für eine jederzeit erweiterbare Menge von Analyseprogrammen dar, denen er die normierten Texte zuführt und deren Analyseergebnisse, die Sekundärdaten, er zur weiteren Verarbeitung als 'Analysedaten' (siehe Bild 2) systemintern ablegt. Neben den Standardanalysen (Textumfang, Informationsgehalt, Redundanz, Verteilung der Wortarten, Type-Token-Ratio) die an jedem Text durchgeführt werden, können zusätzliche Analysen (z.B. Verteilung von Angstthemen) den Wünschen der Benutzer und in Abhängigkeit der verfügbaren Programme entsprechend vorgenommen werden.

Der Datenbankbaustein

Der Datenbankbaustein ist ein Teilsystem zur Verwaltung aller Primär- und Sekundärdaten, also der Informationen über die Texte (B3 und B5 in Bild 2). Er gliedert sich in Programme zur Datendefinition, zur Datenaufnahme und zur Datenauswertung. Die Datendefinition dient der Vereinbarung von Bezeichnungen für die Merkmale zur Beschreibung von Texten und zur Bestimmung der Beziehungen zwischen diesen Merkmalen. Es sind einfache hierarchische und relationale Datenstrukturen möglich. Ein vereinbartes Beschreibungsschema kann jederzeit geändert oder erweitert werden. Die Datenkonsistenz wird dabei überwacht. Die Leistungen dieses Programms werden über eine Benutzerschnittstelle in Form einer Datendefinitionssprache realisiert.

Das Speichern, Löschen und Ändern aller über das Schema beschreibbaren Informationen wird durch die Programme zur Datenaufnahme verwirklicht. Sie kommunizieren ebenfalls über eine Benutzerschnittstelle in Form einer Datenhandhabungssprache. Während der Datenaufnahme wird geprüft, ob die im Schema angegebenen Restriktionen eingehalten

sind.

Anfragen werden durch die Programme zur Datenauswertung bearbeitet. Sie können über eine Benutzerschnittstelle in einer Datenanfragesprache gestellt werden und müssen sich auf das Schema beziehen. Die Ergebnisse einer Anfrage werden dem Benutzer unmittelbar am Bildschirm eingespielt und, falls erwünscht, systemintern an den Textbaustein zur Ausgabe weitergeleitet. In Bild 2 sind diese drei Schnittstellen als 'Benutzereingaben' festgehalten.

Die Erfassung der Sprecherkenndaten (B5 in Bild 2) wird von einem gesonderten maskenorientierten Erfassungsprogramm durchgeführt. Damit ist insbesondere der Schutz personenbezogener Daten in der Praxis leichter zu handhaben. Die Struktur der systemintern abgelegten 'Sprecherkenndaten' (siehe Bild 2) entspricht jedoch der oben erwähnten Datenhandhabungssprache.

Der Monitorbaustein

Das Zusammenspiel aller drei Teilsysteme wird durch den Monitor ermöglicht. Er kommuniziert über eine Benutzerschnittstelle in Form einer Steuersprache mit den TBS-Benutzern und hilft ihnen im Umgang mit dem TBS durch einfache Benutzerführung. Mit der Menutechnik werden ihm die einzelnen Systemleistungen zur freien Wahl am Bildschirm angeboten. Je nach ausgewählter Leistung wird der Dialog fortgesetzt, indem Masken zum Ausfüllen am Bildschirm eingespielt werden oder über eine leicht erlernbare formale Sprache dem System vom Benutzer sein Wunsch mitgeteilt wird.

Gesichtspunkte zur Implementierung

Die Implementierung des TBS ist von dem Wunsch nach Portabilität, von den verfügbaren Rechnern - der Groß-

rechner TR440 und das Mikrocomputersystem TELECOMP - sowie von der stets knappen Manpower gekennzeichnet. Als Implementierungssprache blieb daher als 'kleinster gemeinsamer Nenner' von Groß- und Mikrorechner lediglich FORTRAN IV, was neben den bekannten Nachteilen insbesondere bei der Textverarbeitung wenigstens eine hohe Wahrscheinlichkeit für eine problemlose Portabilität mit sich bringt. Der Datenbankbaustein ist an dem CODASYL-Modell orientiert. Die Datenhandhabungssprachen arbeiten auf einem prädikatenlogischen Ansatz (MERGENTHALER 1976). Die unterschiedlichen Speicher- und Suchprobleme wurden mit einem Unterprogrammsystem realisiert, das auf der Basis der B-Baum-Technik mit beliebig vielen orthogonalen logischen Dateien, jedoch nur einer physischen Datei arbeitet (MERGENTHALER 1979). Hierzu gehören auch einige Routinen zur Verarbeitung von Strings.

Bei der Rechenzeitplanung wurde auf kurze Antwortzeiten im Anfrageteil des TBS Wert gelegt. Das Zusammenstellen von Textkorpora und das Drucken von Texten, was ohnehin im Stapelbetrieb abläuft, ist nicht zeitoptimiert. Ebenfalls verzichtet wurde auf eine Kompression der Texte, da genügend Hintergrundspeicher - notfalls Magnetbänder - zur Verfügung steht und damit Rechenzeit gespart werden kann.

Textbank und Textkorpus

Der optimale Einsatz eines TBS zur Psychotherapieforschung setzt voraus, daß der zu verwaltende Textbestand die zu erwartenden Fragestellungen erfüllen kann. Der Definition einzelner Textkorpora als Untermengen der Textbank kommt daher besondere Bedeutung zu. Für die ULMER TEXTBANK haben sich dabei zwei Schwerpunkte ergeben, die unterschiedlichen Forschungsansätzen, den

> Längsschnittuntersuchungen einerseits und den Querschnittsuntersuchungen andererseits

entsprechen. Die Längsschnittuntersuchungen konzentrieren sich auf Sprachmaterial aus psychoanalytischen Behandlungen und haben die Erforschung des psychoanalytischen Prozesses zum Ziel. Aufgrund der hohen Stundenzahl einer psychoanalytischen Behandlung können nur Transkripte von wenigen verschiedenen Behandlungen erwartet werden, so daß Einzelfallstudien unter dem Aspekt der Sprachveränderung durch den Therapieprozeß im Vordergrund stehen. Therapeuten- und patientenübergreifende Fragestellungen durchgeführt anhand sprachlichem Material aus der Erstinterviewsituation sind Gegenstand der Querschnittsuntersuchungen. Bei den Erstinterviews lassen sich dazu viele verschiedene Patienten mit jeweils einem Gespräch erfassen, so daß sprecherübergreifende Sprachuntersuchungen etwa im Hinblick auf Geschlechtsspezifität, Diagnose usw. möglich werden. Daneben werden Textkorpora geführt, die für spezielle Untersuchungszwecke benötigt werden, wie etwa der Balint-Gruppen-Forschung, der Untersuchung von Visitengesprächen oder der Untersuchung von Familien-Beratungsgesprächen.

Aus den in der ULMER TEXTBANK enthaltenen Texten werden begleitend allgemeine sprachstatistische Eigenschaften der gesprochenen Sprache ermittelt. So wird etwa bei der Aufnahme jedes neuen Textes ein Häufigkeitswörterbuch nebst Konkordanz fortgeschrieben. Für inhaltsanalytische Untersuchungen werden weiterhin allgemeine Bezugsdaten ermittelt, die als Baseline bei speziellen Untersuchungen herangezogen werden können. Zum Vergleich dieser, aus der psychotherapeutischen Situation stammenden Daten mit einer Normalpopulation, wurden in die ULMER TEXTBANK weiterhin 150 Interviews mit Angehörigen bundesdeutscher Haushalte, entstanden im Rahmen einer soziologischen Untersuchung, mit aufgenommen. Ebenfalls für Vergleichszwecke steht eine Sammlung von Interviews mit psychotischen Patienten aus geschlossenen Anstalten zur Verfügung.

Repräsentativität

Die ULMER TEXTBANK stellt in erster Linie eine Basis für die empirische Erforschung des psychoanalytischen Prozesses dar. Fragen der Repräsentativität orientieren sich somit an diesem Forschungsziel. Allerdings sind den hierzu möglichen theoretischen Überlegungen von vornherein praktische Grenzen gesetzt. Die hohe Zahl an Behandlungsstunden und die damit einhergehende Dauer einer Psychoanalyse grenzen den möglichen Spielraum vollkommen ein. Ein Stichprobenumfang von zehn Psychoanalysen ist daher eher willkürlich gewählt. Er läßt allerdings statistische Aussagen zu und ist praktisch zu realisieren. Tatsächlich enthält das Tonbandarchiv unter den 17 teils laufenden, teils schon abgeschlossenen Psychoanalysen rund zehn, deren Übernahme in die Textbank angebracht scheint. In Zahlen ausgedrückt bedeutet dies rund fünf bis zehn Millionen Textwörter verteilt auf 1000 Textexemplare (Stunden). Bei der Auswahl einzelner Behandlungen zur Aufnahme in die Textbank spielen neben den praktischen Problemen die zahlenmäßige Ausgewogenheit der verschiedenen Therapeuten, Diagnosen und Behandlungsdauer eine Rolle. Weitere, besonders im Hinblick auf statistische Auswertungen hin relevante Auswahlkriterien wie die Ausgewogenheit des Geschlechts bei Patienten bzw. Therapeuten sowie der sozialen Schichtzugehörigkeit bei Patienten usw. lassen sich wegen der geringen, überhaupt möglichen Anzahl so gut wie nicht berücksichtigen. Das Ulmer Psychoanalysekorpus wird also immer nur im Hinblick auf einen Forschungszweck als repräsentativ bezeichnet werden können.

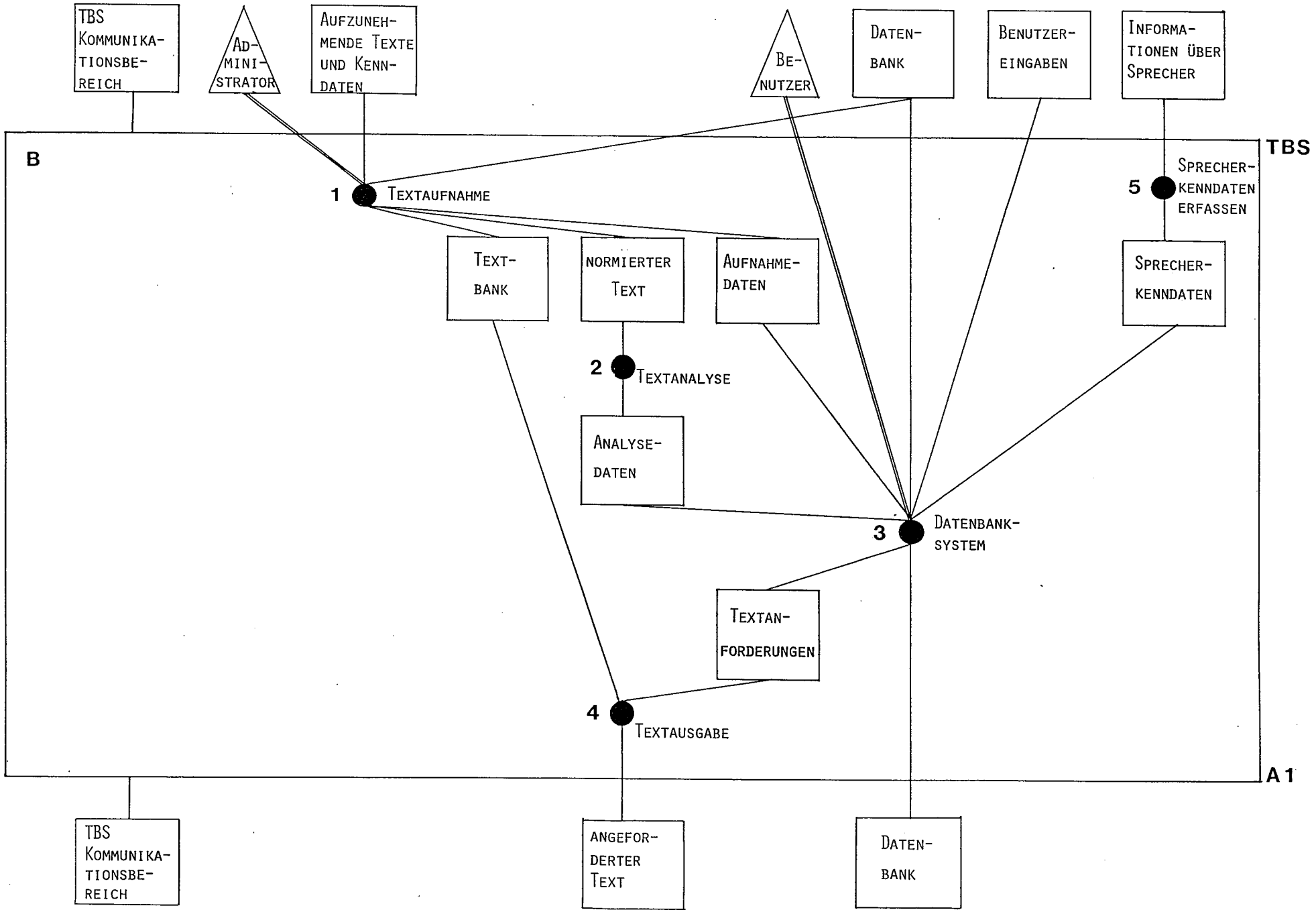
Schutz personenbezogener Daten

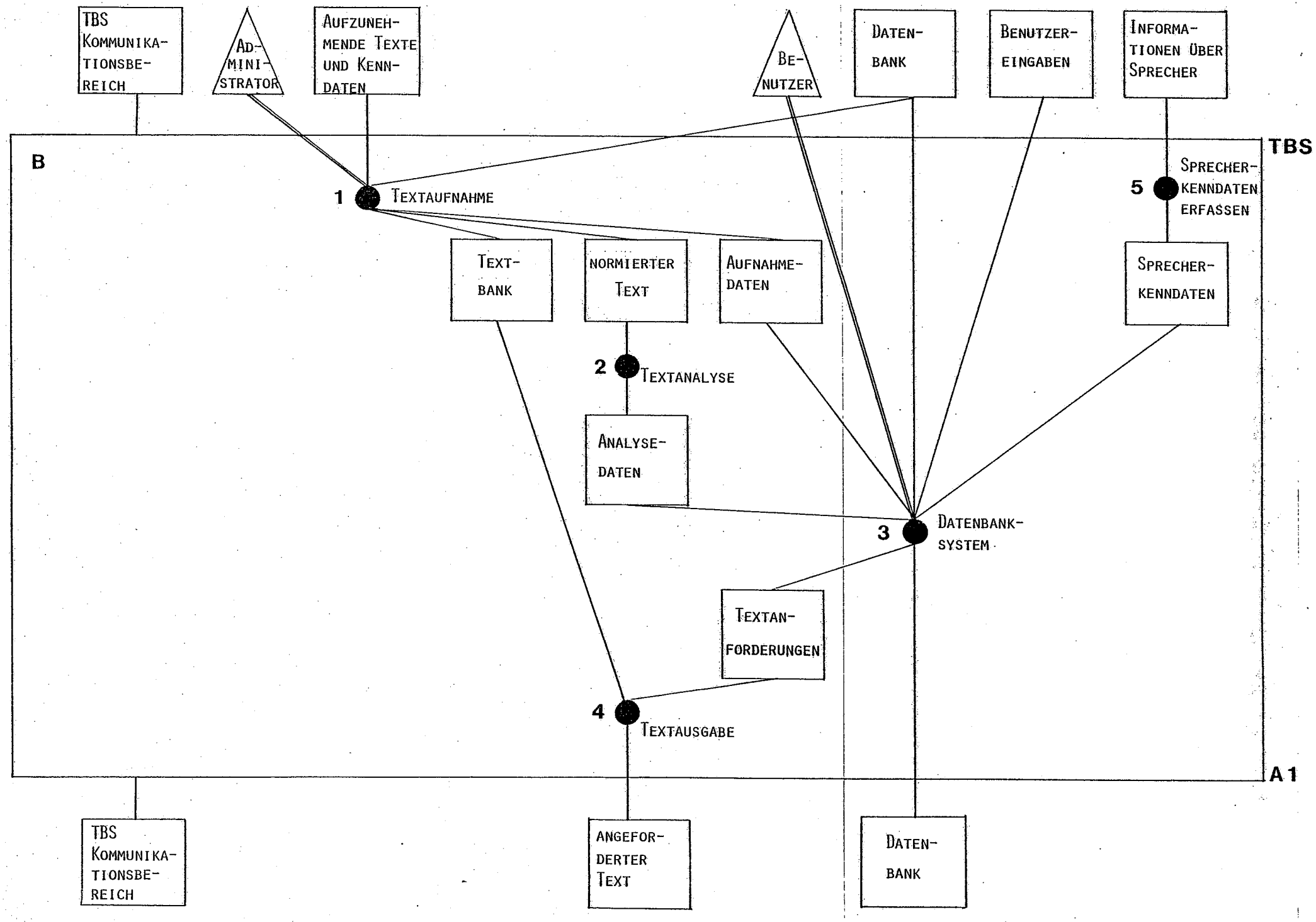
Bei der Aufnahme eines Textes in die Textbank werden die darin enthaltenen Eigennamen, Orts- und Landschaftsbezeichnungen sowie weitere personenbezogene Merkmale durch kryptografische Verfahren (RYSKA und HERDA 1980) verschlüsselt oder durch Pseudonyme ersetzt. Während die

somit faktisch anonymisierten (siehe hierzu Schlörer) Texte auf der Rechenanlage der Universität Ulm weiterverarbeitet werden, verbleiben die Schlüsseldateien, also alle personenbezogene Daten, auf den ausschließlich der ULMER TEXTBANK zur Verfügung stehenden Mikrocomputern. Durch diese verteilte Speicherung sowie durch umfangreiche Zugangs- und Zugriffskontrollen ist die ULMER TEXTBANK weitgehend vor Mißbrauch geschützt. Das mit Arbeiten an der Textbank betraute Personal unterliegt der Schweigepflicht und wurde über die einschlägigen Datenschutzbestimmungen belehrt. Die ULMER TEXTBANK ist im Datenschutzregister des Landes Baden-Württemberg eingetragen.

Verfügbarkeit und Kosten des Systems

Die Konzeption des Textbanksystems ist weitgehend anlagenunabhängig ausgelegt. Die Implementierungssprache FORTRAN IV ist allgemein verfügbar. Sämtliche anlagenspezifischen Komponenten, wie sie insbesondere zur Speicherung der Texte auf Hintergrundspeichern unumgänglich sind, werden in einem Teilbaustein zusammengefaßt und können prinzipiell für jede beliebige Rechenanlage neu erstellt werden. Dies kann jedoch bei der Vielfalt und Unterschiedlichkeit heutiger Computer dennoch mit erheblichen Schwierigkeiten verbunden sein.


TBS KOMMUNIKATIONSBEREICH
ADMINISTRATOR
AUFZUNEHMENDE TEXTE UND KENNDATEN
BENUTZER
DATENBANK
BENUTZEREINGABEN
INFORMATIONEN ÜBER SPRECHER
B
TBS
1 TEXTAUFNAHME
5 SPRECHERKENNDATEN ERFASSEN
TEXTBANK
NORMIERTER TEXT
AUFNAHMEDATEN
SPRECHERKENNDATEN
2 TEXTANALYSE
ANALYSEDATEN
3 DATENBANKSYSTEM
TEXTANFORDERUNGEN
4 TEXTAUSGABE
A1
TBS KOMMUNIKATIONSBEREICH
ANGEFORDERTER TEXT
DATENBANK


TBS KOMMUNIKATIONSBEREICH
ADMINISTRATOR
AUFZUNEHMENDE TEXTE UND KENNDATEN
BENUTZER
DATENBANK
BENUTZEREINGABEN
INFORMATIONEN ÜBER SPRECHER
TBS
B
1 TEXTAUFNAHME
5 SPRECHERKENNDATEN ERFASSEN
TEXTBANK
NORMIERTER TEXT
AUFNAHMEDATEN
SPRECHERKENNDATEN
2 TEXTANALYSE
ANALYSEDATEN
3 DATENBANKSYSTEM
TEXTANFORDERUNGEN
4 TEXTAUSGABE
A1
TBS KOMMUNIKATIONSBEREICH
ANGEFORDERTER TEXT
DATENBANK